

## **Improved Sampling for Queries**

### **Field of the Invention**

5        This invention relates generally to the field of computers, and in particular to improving sampling techniques for more accurate estimations of queries of large databases.

### **Copyright Notice/Permission**

10        A portion of the disclosure of this patent document contains material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever. The following notice applies  
15 to the software and data as described below and in the drawing hereto: Copyright © 2000, Microsoft Corporation, All Rights Reserved.

### **Background**

Decision support applications such as On Line Analytical  
20 Processing(OLAP) and data mining tools for analyzing large databases are gaining popularity. Many databases are growing larger and larger, making it difficult for application programs and computers to search them in their entirety for each request for data, or query, by a user. This is particularly true when the number of users desiring data from the same database is also increasing.

25        Executing such applications on large volumes of data can also be resource intensive. Fortunately though, samples of the results of queries to the database back end can be used to enhance the scalability of application servers without compromising the quality of their analysis. In other words, only portions or samples of the database are used to provide an estimate of results  
30 without scanning the entire database.

Typical decision-support applications use a structured query language (SQL) database system as the backend data warehouse and communicate data retrieval requests via relational SQL queries. On a large database, the cost of

executing such queries against the relational backend can be expensive.

Therefore, the ability of the application servers to work with samples raises the question of whether a database server can compute a sample of answers to queries without paying the cost of executing the entire query. Such functionality  
5 would also enable the database servers to scale efficiently. No matter how many records are in the database, only a statistically relevant sample of them would need to be checked to provide a fairly accurate estimation.

Data mining and statistical analysis involve techniques that can be robust and effective even when supplied with merely a sample of query results.

10 Likewise, OLAP servers that answer queries involving aggregation (performing some function on the results to arrive at an answer, e.g., finding the average) can significantly benefit from the ability to present to the user an approximate aggregate computed from a sample of the subset of the multidimensional data the user is interested in analyzing.

15 It is important to recognize that sampling must be supported on the result of ad-hoc SQL queries, not just on stored relations. There are several reasons for such a requirement. First, applications need to analyze ad-hoc fragments of data, driven by user requirements, e.g., sales data for a certain time period or sales data for a certain product. Next, great scalability benefits may accrue from  
20 speeding-up queries that are expensive to compute in their entirety (complex SQL queries). Such queries typically involve foreign key joins, selections and group by.

Although there has been a wealth of work in databases on supporting sampling, no previous work has carefully examined the practicality of sampling  
25 techniques for ad-hoc queries over arbitrary databases. Approximate query answering has recently received a lot of attention, particularly in the context of Data Warehousing. Since most queries in the warehousing application involve aggregation, there is a need to ensure that approximately answering aggregation queries using sampling provides good results.

30 Sampling performs poorly when the distribution of the aggregated attribute is skewed. In other words, if some values are very high or very low compared to the rest of the values, sampling of the data might not find those "outlier" values, which could greatly affect the result obtained. Such a database

is said to exhibit data skew. Sampling is also hard to use effectively in queries with low selectivity.

The potential effect of outlier values is illustrated by looking at a small database of six records. The values to be aggregated are 1, 2, 1, 4, 3, and 1000.

5 The sampling of these values is based on looking at every other record. The values obtained from this data given the sampling used are 1, 1, and 3. If trying to estimate an average, it is clear that the estimation will be far from the real answer because the value of 1000 was not included in the results. Perhaps the numbers represent sales results from various locations. If trying to estimate total  
10 sales, the estimate will be far to low.

There are many other examples of how outlier values can dramatically affect the results of sampling a database. One example involves 10,000 tuples of which 99% have value 1 in the aggregate column, while the remaining 1% of the tuples are of value 1000. Consider using a uniform random sample of size 100  
15 to estimate the average value of the column over all tuples. It is quite likely that the sample would not include any tuple of value 1000, leading to an estimate of 1 for the average value. On the other hand, if perchance two or more tuples of value 1000 were to be included in the sample, then the estimate of the average value would be more than 20.98. In either case, the estimate would be far from  
20 the true value of the actual average, 10.99. It is only in the case where exactly one tuple of value 1000 is in the sample that a reasonable estimate of the average value would be obtained. But the latter event has probability 0.37, therefore there is high probability that a large error in the estimate would result.

In this example it is observed that a skewed database is characterized by  
25 the existence of certain tuples that are deviant from the rest with respect to their aggregate value. These are referred to as outliers. It should be noted that the presence of these tuples in the wrong proportion in the sample would also lead to large errors in the estimate of a query's result.

There are many examples one could provide to justify that indeed such  
30 skew exists in databases. Consider a database that maintains the population for different countries. The population figures for countries like China and India are surely deviant from the figures for the remaining countries and lead to a skew. Suppose a database maintained nationwide sales figures for a range of items

FOI b7D b7C b7E  
ranging from a particular luxury model of automobile to six-packs of soda pop. While the sales figures for the automobile model would be very small, the sales figures for Coke could range in the millions. One potential solution to the problem involves greatly increasing the sample size. Greatly increasing the sample size results in losing the very benefit that sampling provides. It requires more system resources and takes more time.

There is a need to properly account for such outlier values without significantly affecting the speed and scalability benefits provided by sampling.

Low selectivity of queries can also contribute to significant error in approximation of aggregate values. No single sample of the database can answer all queries with low selectivity with sufficient accuracy. Most queries involve selection conditions or Group-By's. If the selectivity is low, then it dramatically and adversely impacts the accuracy of sampling-based estimation. Selection queries partition the relation into two sub-relations: tuples that satisfy the condition (relevant sub-relation) and those which do not. Sampling in a uniform manner results in the number of tuples that are sampled from the relevant sub-relation being proportional to its size. If the number is low, a very small relevant sample results, leading to large error. The same is true for Group-By queries which partition the relation into numerous sub-relations (tuples that belong to specific groups).

There is also a need for an accurate way to estimate the results of queries, such as an aggregation of queries having selection conditions and Group-By's. There is a need for such a way that does not greatly increase sample sizes, and retains the benefits that sampling provides for databases having data skew and low selectivity.

### Summary of the Invention

Given a query with selection conditions or Group-By's, workload information is used to tune a sample by weighting tuples in the sample. Groups that are accessed more often in the workload are sampled at a higher rate or probability. With each tuple that is included in the sample, the probability is stored. Each aggregate computed over the tuple is then multiplied by the inverse of the probability to normalize it. Using this weighted sampling causes the

sample to contain more tuples that are pertinent to the query, and fewer sampled tuples to be discarded due to not meeting selection criteria where the access patterns of the workload are local, and the workload is a good representation of the actual queries that will be posed in the future.

- 5           Outlier indexes may also be selected using information from the workload for queries with selection conditions or Group-By's. A query with selection or Group-By implicitly defines one or more sub-relations over which aggregates are computed. Workload information is used to weight each sub-relation. The variance in the aggregate column for each sub-relation is
- 10           calculated. An optional pruning retains only sub-relations whose weighted variance is greater than a threshold. Memory is then allocated among all remaining sub-relations in proportion to their weighted variances. An outlier index is then built for each sub-relation. The union of the outlier indexes provides the outlier index for the given query. Queries are then estimated using
- 15           both the weighted outlier index and the weighted sample.

#### **Brief Description of the Drawings**

- Figure 1           is a block diagram of a computer system on which the present invention may be implemented.
- 20   Figure 2           is a flowchart of a sampling process taking outliers into account.
- Figure 3           is a flowchart of an outlier identification algorithm.
- Figure 4           is a graphic representation of windows used in the algorithm of Figure 3.
- Figure 5           is a flowchart of the use workload information in the context of
- 25           weighted sampling.
- Figure 6           is a flowchart of the collection of weights for multiple sub-relations given a query with low selectivity.
- Figure 7           is a flowchart of the generation of an outlier index for multiple sub-relations given a query with low selectivity.

30

#### **Detailed Description**

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings that form a part

hereof, and in which is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that logical, mechanical, electrical and other changes may be made without departing from the spirit or scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

The detailed description is divided into multiple sections. A first section describes the operation of a computer system that implements the current invention. This is followed by a description of the problems of sampling when the data being sampled exhibits data skew or low selectivity, and a high level description of how aggregate queries are estimated using a combination of outlier identification and sampling. Further detail regarding the identification of outlier data, the sampling techniques and how data from outliers and sampling are combined is then described. One further section provides detail regarding the exploitation of workload information for queries with selection and group by. This is followed by a conclusion which describes some potential benefits and describes further alternative embodiments.

20

#### Hardware and Operating Environment

Figure 1 provides a brief, general description of a suitable computing environment in which the invention may be implemented. The invention will hereinafter be described in the general context of computer-executable program modules containing instructions executed by a personal computer (PC). Program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Those skilled in the art will appreciate that the invention may be practiced with other computer-system configurations, including hand-held devices, multiprocessor systems, microprocessor-based programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like which have multimedia capabilities. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices linked

30

through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

Figure 1 shows a general-purpose computing device in the form of a conventional personal computer 20, which includes processing unit 21, system memory 22, and system bus 23 that couples the system memory and other system components to processing unit 21. System bus 23 may be any of several types, including a memory bus or memory controller, a peripheral bus, and a local bus, and may use any of a variety of bus structures. System memory 22 includes read-only memory (ROM) 24 and random-access memory (RAM) 25. A basic input/output system (BIOS) 26, stored in ROM 24, contains the basic routines that transfer information between components of personal computer 20. BIOS 26 also contains start-up routines for the system. Personal computer 20 further includes hard disk drive 27 for reading from and writing to a hard disk (not shown), magnetic disk drive 28 for reading from and writing to a removable magnetic disk 29, and optical disk drive 30 for reading from and writing to a removable optical disk 31 such as a CD-ROM or other optical medium. Hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to system bus 23 by a hard-disk drive interface 32, a magnetic-disk drive interface 33, and an optical-drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for personal computer 20. Although the exemplary environment described herein employs a hard disk, a removable magnetic disk 29 and a removable optical disk 31, those skilled in the art will appreciate that other types of computer-readable media which can store data accessible by a computer may also be used in the exemplary operating environment. Such media may include magnetic cassettes, flash-memory cards, digital versatile disks, Bernoulli cartridges, RAMs, ROMs, and the like.

Program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24 and RAM 25. Program modules may include operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into personal

computer 20 through input devices such as a keyboard 40 and a pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial-port interface 46 coupled to system bus 23; but they may be connected through other interfaces not shown in Figure 1, such as a parallel port, a game port, or a universal serial bus (USB). A monitor 47 or other display device also connects to system bus 23 via an interface such as a video adapter 48. In addition to the monitor, personal computers typically include other peripheral output devices (not shown) such as speakers and printers.

Personal computer 20 may operate in a networked environment using logical connections to one or more remote computers such as remote computer 49. Remote computer 49 may be another personal computer, a server, a router, a network PC, a peer device, or other common network node. It typically includes many or all of the components described above in connection with personal computer 20; however, only a storage device 50 is illustrated in Figure 1. The logical connections depicted in Figure 1 include local-area network (LAN) 51 and a wide-area network (WAN) 52. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When placed in a LAN networking environment, PC 20 connects to local network 51 through a network interface or adapter 53. When used in a WAN networking environment such as the Internet, PC 20 typically includes modem 54 or other means for establishing communications over network 52. Modem 54 may be internal or external to PC 20, and connects to system bus 23 via serial-port interface 46. In a networked environment, program modules, such as those comprising Microsoft® Word which are depicted as residing within 20 or portions thereof may be stored in remote storage device 50. Of course, the network connections shown are illustrative, and other means of establishing a communications link between the computers may be substituted.

Software may be designed using many different methods, including object oriented programming methods. C++ and Java are two examples of common object oriented computer programming languages that provide



functionality associated with object oriented programming. Object oriented programming methods provide a means to encapsulate data members (variables) and member functions (methods) that operate on that data into a single entity called a class. Object oriented programming methods also provide a means to  
5 create new classes based on existing classes.

An object is an instance of a class. The data members of an object are attributes that are stored inside the computer memory, and the methods are executable computer code that act upon this data, along with potentially providing other services. The notion of an object is exploited in the present  
10 invention in that certain aspects of the invention are implemented as objects in one embodiment.

An interface is a group of related functions that are organized into a named unit. Each interface may be uniquely identified by some identifier. Interfaces have no instantiation, that is, an interface is a definition only without  
15 the executable code needed to implement the methods which are specified by the interface. An object may support an interface by providing executable code for the methods specified by the interface. The executable code supplied by the object must comply with the definitions specified by the interface. The object may also provide additional methods. Those skilled in the art will recognize that  
20 interfaces are not limited to use in or by an object oriented programming environment.

#### Effects of Skew and Low Selectivity

25 The effects of skew and low selectivity can adversely affect the accuracy of sampling-based estimations.

The following example demonstrates the adverse impact of skew on the applicability of sampling.

Example 1. Suppose there are 10,000 tuples of which 99% have value 1  
30 in the aggregate column, while the remaining 1% of the tuples have value 1000. Consider using a uniform random sample of size 100 to estimate the average value of the column over all tuples. It is quite likely that the sample would not include any tuple of value 1000, leading to an estimate of 1 for the average

value. On the other hand, if perchance two or more tuples of value 1000 were to be included in the sample, then the estimate of the average value would be more than 20.98. In either case, the estimate would be far from the true value of the average which is 10.99. It is only in the case where there is exactly one tuple of value 1000 in the sample that we would obtain a reasonable estimate of the average value. But the latter event has probability only 0.37. Therefore, with high probability a large error in the estimate would result.

In this example it is observed that a skewed database is characterized by the existence of certain tuples that are deviant from the rest with respect to their contribution to the aggregate value. These tuples are referred to as "outliers." The presence of outliers causes a large "variation" in the aggregate values which is directly proportional to the standard error in sampling as demonstrated by the following theorem. The theorem quantifies the error for estimating aggregate queries that calculate sum or average of values defined over the tuples. The value associated with a tuple could be a function of one or more of its attributes (columns). For example, if the query is `select sum (quantity*price)` from `lineitem` then the function  $f(t)$  over the tuples is  $f(t) = t.quantity * t.price$ .

**Theorem 1.** Consider a relation of size  $N$  and let  $\{y_1, y_2, \dots, y_N\}$  be the set of values associated with the tuples in the relation. Let  $U$  be a uniform random sample of the  $y_i$ 's of size  $n$ . Then the sample mean  $\bar{y} = 1/n \sum_{y_i \in U} y_i$  is an unbiased estimator of the actual mean  $\bar{Y} = 1/N \sum_{i=1}^N y_i$  with a standard error of

$$\epsilon = \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} \quad (1)$$

where

$$S = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N - 1}}$$

If there are outliers in the data then the quantity  $S$  is very large, in which case the sample size  $n$  should be increased to meet the error bounds. For instance, consider two sets of values each containing  $N$  elements. Let  $S$  and  $2S$  be the variance of the two sets. If a sample size of  $n$  is sufficient to ensure error bounds for the first set, a sample size  $4n$  is required to provide the same

guarantees for the second set. In general, the sample size would have to increase quadratically with the variance thereby hampering the efficiency gained from sampling. Earlier works assume that the aggregate attributes are not skewed.

Since most queries involve selection conditions and/or Group-By's, it is  
5 important to study their interaction with sampling. If the selectivity of a query is low, then it dramatically and adversely impacts the accuracy of sampling-based estimation. A selection query partitions the relation into two sub-relations: tuples that satisfy the condition (relevant sub-relation) and those which do not. With sampling uniformly from the relation, the number of tuples that are  
10 sampled from the relevant sub-relation is proportional to its size. If this number is small (low selectivity), very small relevant sample size is obtained, leading to large error. Consider the standard error ( $\epsilon$ ) in uniform sampling as given by Theorem (1). If  $n$  is the sample size for the entire relation and  $q$  is the selectivity of the query, then the expected number of samples that satisfy the selection  
15 condition is given by  $qn$ . Since the error is inversely proportional to  $\sqrt{qn}$ , a low selectivity  $q$  will give rise to large error due to sampling. The same is true for Group-By queries which partition the relation into numerous sub-relations (tuples that belong to a specific group). The main point to note is that for uniform sampling to perform well, the importance of a sub-relation should be  
20 reflected by its size, which is not the case in general.

#### Use of Outlier Indexes to Approximate Aggregate Queries

As indicated above, datasets may contain outlier values, which are values  
25 that occur with low frequency, but have a high contribution to an aggregate of the data being computed. Uniform random sampling is not sufficient to provide an accurate estimate of the aggregate when the data being sampled contains such outliers. An improved technique accounts for such outliers in combination with sampling of the data. A high level simplified flowchart of the steps involved is  
30 shown in Figure 2. The functions described in this and other flowcharts may be represented in a computer program stored on a computer readable medium and performed on the computer system of Figure 1. The computer readable medium

may also comprise carrier waves for transmitting the program between computers and various storage medium.

In Figure 2, processing starts at 210. A query is received from a user at 210 and outliers are identified from the data to which the query is directed at 220, and as further described with respect to Figures 3 and 4. The outliers identified are then indexed at 225. It should be noted that the identification of outliers and indexing of them is normally performed prior to receipt of queries, but may also be performed prior to the query being received. This greatly speeds up the processing of the query, which in many cases is real time, with a user waiting for results.

Following creation of the index of outliers at 225, the outliers are then subjected to the query and aggregated at 230. This provides an outlier aggregate, such as a sum of all the outliers meeting the query criteria. A sample is obtained at 232. It may be obtained in many different ways, such as during a preprocessing phase prior to execution of the query. Next, the outliers are removed from the data at 235. Such data is then sampled, subjected to the query, and aggregated. The result of this is extrapolated based on the frequency of the sampling and the total number of non-outliers to provide an estimated aggregate. The estimated sum is combined with the outlier aggregate to provide an estimate of the aggregate of the entire dataset at 240, and control is returned at 250.

The overall process of Figure 2 is now described in further detail. Given a query  $Q$  which aggregates over column  $C$  of relation  $R$ , an existing "outlier index" for  $C$  is used along with a uniform sample of  $R$  to approximately answer  $Q$ . An "outlier index"  $R_O$  may be thought of as a sub-relation of the original relation  $R$ .

The table  $R$  is partitioned into two subtables  $R_O$  (outliers) and  $R_{NO}$  (non-outliers), i.e.,  $R_O$  is the set of tuples that belong to the outlier index. An aggregation query  $Q$  can be considered as the "union" of two sub-queries, the first of which is  $Q$  applied to the  $R_O$ , while the second is  $Q$  applied to the  $R_{NO}$ . This leads to the following scheme for approximately answering an aggregate query for a given choice of the outliers  $R_O$ . To illustrate the scheme, the example query – *select sum(sales) from lineitem* is used.



A simplified overview of the manner in which outlier values are selected is shown in Figure 3. Starting at 310, a threshold,  $L$ , for the number of tuples or rows of a dataset are determined. This threshold may be determined based on the availability of system resources that can be dedicated to improving the accuracy of an estimated aggregation, and can also be based on a desired accuracy of the estimation. The threshold is basically the number of outlier values which can be taken into account, and may be set by a database system administrator. Further detail is provided in the alternative embodiments section.

Given the threshold, values in a column being aggregated are sorted at 320.  $L$  corresponds to the number of values, which may be equal to the number of tuples in the database if there are no selection conditions. The values are sorted in ascending or descending order using any available sorting algorithm. Once the values have been sorted, an index of outlier values is created. A sliding window is employed to identify sets of outliers. In other words, the window is of size  $N-L$ , and is slid along the column of values, starting at the first value in block 330. At this point, the first  $N-L$  values are situated within the window, and the last  $L$  values are outside of the window. A variance of such values is calculated at 340. The window is then moved down one value in the sorted list, such that the first value is no longer in the window. Since the window is a constant size in one embodiment,  $L-1$  values at the high end of the list are now excluded from the window. As the window is slid one value at a time, the variance of the values within each window of size  $N-L$  is determined at 350. The variance is determined in one embodiment by maintaining it for a dynamic set of numbers subject to insertions and deletions. Moving the window one value results in both an insertion and deletion rather than a recalculation of the variance using all the numbers in the set or window.

The index is created by identifying the window having the lowest variance, and including all values outside of the window in the outlier index at 360. Several variations of the above are possible, including sliding the window more than one value at a time. In one embodiment, the window moves multiple values between calculations of the variance. In this embodiment, once the low variance window is determined, the window may be repositioned one value at a time about such low variance window and the variance at each window

calculated. Again the window with the low variance is selected and outliers thus identified. In still further embodiments, it is recognized that once a lowest variance is encountered, and the variance keeps increasing, one may stop sliding the window and select the window with the lowest variance already calculated.

5        A simple example of selecting the outlier values and finding the estimated aggregate is described with respect to Figure 4. Twelve values are shown, corresponding to twelve tuples in a database. The values are shown four times, in columns 420, 425, 430 and 435. A window is also shown in each column, identifying nine entries in each column. The windows are slid down  
10 each column, such that the first window 440 includes the first nine values. The second window 445 covers values two through 10, the third window covers values three through eleven, and the fourth window covers values four through twelve. Each window is shown with a variance corresponding to the variance of the values in the window. The variance of column 425 is lowest, and defines the  
15 values to be sampled. In this example, every third value is included in the sample, and there is memory allocated to three outliers.

Further, in this example, the sum of the column of values is the desired aggregate. The actual sum is easily seen to be two hundred twenty two. To estimate the sum in accordance with the invention, the sum of the outliers,  
20 corresponding to values one, forty, and eighty is one hundred twenty one. Every third sample of the values in the window 445, include the values ten, fourteen and fifteen. The sum of these samples must be multiplied by the inverse of the sample rate, or three, yielding a sum of one hundred seventeen, for a total estimated sum of two hundred thirty eight, as shown at 465. Estimated sums for  
25 the other window positions are also shown at 460, 470 and 480. It should be noted that this example is only for illustration of the operation of the invention, and is not actually a good candidate for estimation, since the total number of values is extremely small. In fact, column 460 provides a slightly better estimate, even though the variance of the data in the window was slightly higher  
30 than that of window 445. However, the variance of windows 450 and 455 was significantly higher, as were the estimated sums using such windows.

#### Detail of Identifying Outlier Values

Further detail is now provided with respect to identifying the outlier values and constructing outlier indexes.

This section describes further detail in selection of outliers where the aggregation is a sum. Since the relative error in estimating the average is the same as that for sum, the choice of outliers does not change when the aggregation is average.

There are two points to be observed: (1) the query error is due to the error in estimating the aggregate of the non-outliers from their sample and (2) there is an extra cost (above that incurred by pure sampling) due to the need for maintaining and accessing the outlier index. Therefore, outlier set  $R_O$  is selected to minimize the approximation error for a class of queries subject to the constraint that  $R_O$  contains at most  $\tau$  tuples from the relation  $R$ . The following definition identifies an optimal choice of the outlier set  $R_O$ .

**Definition 1.** For any sub-relation  $R' \subset R$ , let  $\epsilon(R')$  be the standard error in estimating  $\sum_{t \in R'} t.C$  using random sampling followed by extrapolation. An optimal outlier set  $R_O(R, C, \tau)$  for a column  $C$  in a relation  $R$  with a threshold  $\tau$  is defined as a sub-relation  $R_O \subset R$  such that

$$|R_O| \leq \tau, \text{ and}$$

$$\epsilon(R_O) = \min_{R' \subset R, |R'| \leq \tau} \{\epsilon(R')\}.$$

Essentially the outlier index is defined as an optimal sub-relation  $R_O$  that leads to the minimum possible sampling error, subject to the constraint that  $R_O$  has at most  $\tau$  tuples in the relation  $R$ . The sampling error is the error in estimating the aggregate value over the tuples not included in  $R_O$  using the standard sample-and-extrapolate strategy. The error is directly proportional to

the variance  $S$  given by  $S = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}}$ . Let  $S(R')$  be the variance for any

sub-relation  $R' \subset R$ . Then an outlier index  $R_O$  is a sub-relation of size at most  $\tau$  such that the complement  $R \setminus R_O$  sub-relation has minimum variance  $S(R \setminus R_O)$ . The following theorem assists in choosing such a sub-relation efficiently.

**Theorem 2.** For any multiset  $A = (y_1, y_2, \dots, y_n)$  let the variance  $S(A)$  be

defined as  $S = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}}$  where  $\bar{Y} = 1/N \sum_{i=1}^N y_i$  is the mean of the



multiset. Consider a multiset  $R = \{y_1, y_2, \dots, y_N\}$  where the  $y_i$ 's are in sorted order. Let  $R_O \subset R$  be the subset such that

$$|R_O| \leq \tau, \text{ and}$$

$$S(R \setminus R_O) = \min_{R' \subset R, |R'| \leq \tau} \{S(R \setminus R')\}.$$

5 then  $R_O = \{y_1, y_2, \dots, y_{\tau'}, y_{N-\tau+\tau'+1}, \dots, y_N\}$  for some  $0 \leq \tau' \leq \tau$ .

The theorem states that the subset that minimizes the variance over the remaining set consists of the leftmost  $\tau'$  elements (for some  $0 \leq \tau' \leq \tau$ ) and the rightmost  $\tau - \tau'$  elements from the original multiset, when the elements are arranged in a sorted order. Thus, the selection of an outlier index simplifies to  
10 finding the value  $\tau'$ . This gives rise to the following algorithm for outlier index selection, which follows the flowchart of Fig. 3.

Algorithm Outlier-Index ( $R, C, \tau$ ):

1. Read the values in column  $C$  of the relation  $R$ . Let  $y_1, y_2, \dots, y_N$  be the sorted order of the values appearing in column  $C$ . Each value corresponds to a  
15 tuple (320).
2. For  $i=1$  to  $\tau$ , compute  $E(i) = S(\{y_i, y_{i+1}, \dots, y_{N-\tau+i-1}\})$ , where  $S$  is the variance of the multiset as defined in Theorem 2 (350).
3. Let  $\tau'$  be the index where  $E(i)$  takes its minimum value. Then, the outlier  
index is the tuples that correspond to the set of values  $\{y_1, y_2, \dots, y_{\tau'-1}, y_{N-\tau+\tau'},$   
20  $y_{N-\tau+\tau'+1}, \dots, y_N\}$  (360).

The efficiency of the algorithm crucially depends on the ability to compute variance efficiently. We show how this can be done in linear time, once the values in column  $C$  have been sorted. It is well known that quantities such as sum, mean, variance, etc. can be efficiently maintained for a dynamic set  
25 of numbers subject to insertions and deletions. In particular, each insertion or deletion can be handled in  $O(1)$  time. In implementing Step 2, we start by scanning the data to compute  $E(1)$ . We also make memory-resident copies of the first  $\tau$  and the last  $\tau$  values of  $(y_1, y_2, \dots, y_N)$ . (If they cannot fit into memory we make copies on disk, however  $\tau$  is usually small). After this we do not have  
30 to scan the data any more. This is because  $E(i+1)$  can be incrementally computed from  $E(i)$  since the value to be deleted ( $y_i$ ) and the value to be inserted

( $y_{N-r+i}$ ) can be looked up from the copies. Thus, the running time of the algorithm is dominated by the sorting in Step 1.

Multiple aggregates may be handled by maintaining an outlier index for each aggregate. However, outlier indexes for different aggregates are likely correlated, and as a result have many tuples in common (a tuple that is outlier for a certain column or aggregate may also be an outlier for other aggregates). Hence, an optimization heuristic is to maintain the union of these indexes. A more precise analysis partitions the available memory space among outlier indexes for multiple aggregates in a judicious manner that takes into account their relative importance, e.g., by exploiting workload information.

Given sufficient space to store  $T$  tuples, the allocation of memory between samples and outlier index should be done in a manner that minimizes error.

Let  $S(n)$  denote the variance in the non-outliers for an optimal outlier index of size  $n$ . If we allocate the space such that we have  $n$  tuples in the outlier index and  $T - n$  tuples in the sample, then the error as given by equation 1 is proportional to  $S(n)/\sqrt{T - n}$ . Since the function  $S(n)$  does not have a closed-form formula and depends on the data distribution, identifying an optimal allocation would require evaluating  $S(n)$  for a set of evenly-spaced values of  $n$  from 0 to  $T$ ; finding the value of  $n$  ( $n_o$ ) for which  $S(n)/\sqrt{T - n}$  is minimized and distributing the memory as  $n_o$  tuples in the outlier index and  $T - n_o$  tuples in the sample.

Extensions to Other Aggregates. We have described a procedure for selecting outliers for the case of *sum* and *avg*. Suppose we want to aggregate (sum or average of)  $f(t)$ , where  $f$  is a real valued function defined over the tuples. Then, we can use the function values instead of  $y_i$ 's in our algorithm and can determine the sub-relation that minimizes the variance over the remaining set. One should note that  $f(t) = 1$  for the count aggregate and in this case there is no variance in the values. In other words, outlier indexing is not required for count aggregates and uniform sampling works well. It is easy to see that if the function is linear in  $y_i$ 's (i.e.,  $f(t) = ay_i + b$ ), then the same set of tuples that minimize the variance for  $y_i$ 's minimize the variance for  $f(t)$ 's and hence the

outlier index defined for  $y_i$ 's can be used. It is well known that uniform sampling works well for estimating the median. On the other hand sampling does not work well for *min* (and *max*) queries since intuitively, the min value may be much smaller than the next smallest value, and therefore our estimate of *min* can be arbitrarily bad unless we happened to obtain the *min* value in the sample itself. However, we can use sampling to return a value that is not too far away from the *min* or *max* in terms of the rank ordering, i.e., we can use sampling to return a value which, while not the minimum, would be guaranteed to be among the  $k$  smallest values for some small  $k$ ; of course, the  $k$ th smallest value may be significantly different from the smallest value.

### Exploiting Workload Information for Sampling

The sliding window approach was described above for use in estimating aggregation queries where data skew existed. In this section, workload information is leveraged for estimating aggregation queries having low selectivity. The workload information is used both in obtaining the sample, and in creating the outlier index. By tuning the sample and outlier index to a representative workload (i.e., set of queries), the queries can be estimated more effectively.

The use of workload information for sampling and outlier indexing involves the following steps as shown in the flowchart of Figure 5:

1. Workload Collection 510: Obtain a workload representative of the queries posed against the database. Modern database systems provide tools to log queries posed against the server (e.g., the Profiler component of Microsoft SQL Server).
2. Trace Query Patterns 520: The workload can be analyzed to obtain parsed information, e.g., the set of selection conditions that are posed.
3. Trace Tuple Usage 530: The execution of the workload reveals additional information on usage of specific tuples, e.g., frequency of access to each tuple, the number of queries in the workload for which it passes the selection condition of the query. Since tracking this information at the level of tuples can be expensive, it can be kept at coarser granularity, e.g., on a page-level. Alternatively, batching of updates can be used to lower this overhead.

4. Weighted Sampling 540: Perform sampling by taking into account weights of the tuples (from 530).
5. Weighted Outlier Indexing 550: Identify an outlier index that captures outliers in popular queries using query pattern analysis (from 540).

5        In principle, the precomputed uniform sample for the entire relation can be used to derive a sample for each sub-relation created by a selection condition or a group-by. However, unlike the case where aggregation is done over the entire relation in which there was exact control on the sample size, in the current case direct sampling is not done from the sub-relation. Instead, sampling is done  
 10       from the entire relation and then the selection filter is applied to the sampled tuples. If  $q$  is the selectivity, for a sample of size  $n$ , the effective sample for the sub-relation would be of *expected* size  $nq$  but may differ from this expectation. To guarantee that with high probability at least  $nq$  tuples are sampled from the sub-relation, slightly more than  $n$  tuples are sampled from the relation. In the  
 15       following discussion this notion of “slightly” is quantified. Note that the discussion is valid for the case of group-by queries too, where the selectivity for a group is defined as the ratio  $N_g/N$ , where  $N_g$  is the number of tuples that belong to the group and  $N$  is the number of tuples in the relation.

      Consider a group that has very few tuples, say  $N_g$ . Suppose each tuple  
 20       with probability  $p$  is sampled, leading to an expected sample size of  $N_gp$ . From Chernoff bounds, it is guaranteed that the actual number of tuples that are sampled from the group is at least  $(1 - \delta)N_gp$  (the expected value is  $N_gp$ ) with confidence at least  $1 - \gamma$  from some  $0 < \gamma, \delta < 1$ , provided  $N_gp$  is sufficiently large. Since the sampling fraction  $p$  is fixed, this requires that  $N_g$  be large.  
 25       Therefore, it cannot be guaranteed that the small groups are represented sufficiently in the sample. Hence, only guarantees for the big groups are provided. Workload information can also be leveraged for the case of small groups as described below. If the selection (similar results are valid for group-bys) condition is independent of the aggregate column and if the selectivity is  
 30       not too small, then uniform sampling works well, otherwise weighted sampling is used.

### Exploiting Weighted Samples

With weighted sampling, sampling is performed more from groups which though small in size, are important, i.e., have high usage. Usage of a database is typically characterized by considerable locality in the access pattern. Queries  
5 against the database access certain parts of the data more than others. Normalizing the aggregate when weights are associated with tuples is more complex since there is no single sampling fraction. With each tuple that is included in the sample 610 in Figure 6, the probability with which it was accepted in the sample is associated or stored at 620. The inverse of this  
10 probability is the multiplication factor associated with the tuple. This essentially says that if a tuple was included with probability  $p$  then it is a representative of  $1/p$  similar tuples. Each aggregate computed over this tuple is multiplied by this multiplication factor at 630. In the (degenerate) case of uniform sampling since the probability is the same for each tuple the multiplication factor is the same for  
15 all the tuples.

Weighted sampling works well if (a) the access pattern of queries is local (most of the queries access a small part of the relation) and (b) the workload is a good representation of the actual queries which will be asked in future. To quantify the benefit of workload-driven weighted sampling, let  $G$  be the set of  
20 tuples accessed by a query  $Q$ . If the query  $Q$  is such that the average weight of a tuple in  $G$  is significantly higher than the overall average weight, then the effective sampling fraction is much higher for this query as compared to the overall sampling rate. This is quantified in the following lemma.

*Consider a relation  $R$  (fact table). Let  $Q$  denote the subset of tuples from  
25  $R$  that are accessed (satisfy the selection condition) by some query. Let  $w_Q$  denote the average weight of a tuple in  $Q$ , and  $w$  denote the average weight of all tuples in  $R$ . Using weighted sampling with an overall sampling fraction  $p$ , the effective sampling fraction for the query equals  $pw_Q/w$ .*

Thus, weighted sampling gives better accuracy for queries which access  
30 tuples that have higher weights (high  $w_Q/w$  ratio). As a result, better accuracy is obtained through effectively doing more sampling although the execution time remains the same since the sample size for the fact table is the same as in the case of uniform sampling. In other words the sample contains more tuples that

are pertinent to the query and less of the sampled tuples are thrown away (do not pass the selection condition).

### Selecting Outlier Indexes According to Workload

5       A query with selection or group-by implicitly defines one or more sub-relations over which aggregates are computed. The set of outliers for the sub-relation produced by a selection condition may be completely different from the outliers for the entire relation (global outliers). In fact, they may not even belong to the set of global outliers. Ideally, in such situations, we would like to  
10   have an outlier index for each possible sub-relation in order to answer aggregate queries pertaining to it, but that would be prohibitively expensive.

      Consider the case where the selection condition (or group-by) is *independent* of the column that is being aggregated. In that case, one would expect that tuples that are outliers in the sub-relation(s) created by the selection  
15   or the group-by are outliers for the entire relation too and are a part of the global outlier index. Therefore, applying the selection condition to the global outlier index will yield an outlier index for the sub-relation. However, the independence assumption need not be valid in many situations. In fact, the independence assumption is less likely to be valid whenever the sub-relation size  
20   is small, e.g., when a condition or a group has low selectivity, or a range query on the aggregate column itself. We outline an approach that exploits workload information to define an outlier index that can cope in the absence of the independence assumption.

      If the independence assumption does not hold, we are left with no choice  
25   but to index the outliers for each sub-relation and consolidate everything into a single global index. Since we cannot afford to make the outlier index too big, we have to ration the available space between different indexes corresponding to sub-relations. Let  $\{R_1, R_2, \dots, R_q\}$  be the (implicit) sub-relations corresponding to the different selection conditions and group-by's. If there is no selection or  
30   group-by, then the corresponding sub-relation is the entire relation. Let  $\{w_1, w_2, \dots, w_q\}$  be the weights (number of queries that queried the sub-relation) associated with these sub-relations for the given workload. Let  $T$  be the total

available memory for building an outlier index. The problem is to ration  $T$  among the sub-relations for building indexes.

In this section we see how we can do this allocation among different sub-relations efficiently. It should be noted that the scheme that we present is based on heuristics.

To allocate the memory among the different sub-relations we propose the following steps as shown in Figure 7:

1. Collect information about the different sub-relations  $\{R_1, R_2, \dots, R_q\}$  and their weights  $\{w_1, w_2, \dots, w_q\}$  using the workload (710).
- 10 2. For each sub-relation  $R_i$  calculate the variance  $S(R_i)$  in the aggregate column for that sub-relation. The error corresponding to this sub-relation is proportional to  $S(R_i)$  (720).
3. Retain only those sub-relations for which the weighted variance  $w_i S(R_i)$  is greater than a certain user-specified threshold. This step prunes out or  
15 eliminates those sub-relations for which the data distribution does not warrant an outlier index or those that are not queried often. In other words, they have a weighted variance below the threshold (730).
4. Allocate the given memory among all remaining sub-relations in proportion to their weighted variances ( $w_i S(R_i)$ ) (740). Using this memory bound, build  
20 an outlier index for each sub-relation (750) and take the union of the outlier indexes over all sub-relations for which outlier indexes were built (760).

Several improvements to the basic scheme presented above are possible. For example, under certain circumstances it is possible to exploit overlap among sub-relations. However, the basic scheme suffers from the complication that the  
25 available memory allocated to outlier indexes may not be fully exploited. This would require having to repeat Steps 3 and 4 with lower user-specified threshold. In an alternative scheme, where it is possible to globally order the outliers, from the most significant to the least, and index them in this order until the available memory is fully used up.

30 Finally, we note that to build a robust system, we need a mechanism to decide for a given query whether our samples and outlier indexes are good enough (i.e., give acceptable error) for answering it. We suggest the following mechanism: We use the sample of the non-outliers to get an estimate of the

variance in the aggregate values over the non-outliers in the sub-relation that is queried. Using this estimate and the number of samples of the non-outliers we can estimate the error using equation in Theorem 1. This error estimate can be used to decide, perhaps with user input, whether we should use the combination of samples and outlier index to answer the query or if we should answer it exactly using the data itself.

The aggregation of the outlier index is then combined with the sample to arrive at the overall estimate of the query as above.

### Conclusion

Estimating aggregation queries is done by first identifying outlier values, aggregating the outlier values, and then sampling the remaining data after pruning the outlier values. The sampled data is extrapolated and added to the aggregated outlier values to provide an estimate for each aggregation query.

The use of outlier values combined with sampling provides the ability to use sampling-based estimations over data having skew in aggregate values, low selectivity or ad-hoc queries and where there are small groups in group-by queries with better accuracy, and without losing the benefits of sampling in reducing system resource requirements.

For queries having low selectivity or group by, weighted sampling is used, and may be combined with an outlier index based on weights of sub-relations. Sampling is performed on the entire relation, and then the aggregates are multiplied by the inverse of the probabilities associated with each sub-relation. The variance of each sub-relation is used to prune sub-relations, and the remaining sub-relations are used to calculate an outlier index for each one. The outlier indexes are then subjected to a union, to form a single outlier index. The aggregate from this single outlier index is then combined with the aggregate from the weighted sample to provide a final estimate of the aggregation.

This application is intended to cover any adaptations or variations of the present invention. It is manifestly intended that this invention be limited only by the claims and equivalents thereof. SQL was used to query the data and create indexes and sampling. Other forms of relational database query language and



various programming languages could be used as known to those of skill in the art.

0975904-01394  
FOOT-40865260